DD2420 - Tutorial 8

Magnus Pierrau

February 2020

1 Background

We want to find the best distribution $q(\boldsymbol{\theta})$ to estimate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ by an iterative scheme which maximizes the *Evidence Lower Bound (ELBO)* of the posterior distribution of the data generated from a Bayesian Gaussian Mixture model, given the latent parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{c})$. Here $\boldsymbol{\mu}$ is a $k \times p$ matrix with column entries $\boldsymbol{\mu}_k$ indicating the cluster centre for the kth cluster, $\boldsymbol{\sigma}^2$ a diagonal $K \times K$ matrix with entry σ^2_{kk} the variance for the kth cluster (all samples are assumed to be independent), and $\boldsymbol{c}$ being a $k \times N$ matrix of indicator vectors, with all entries being either 0 or 1, indicating which cluster sample $\boldsymbol{x}_i$ belongs to. If $c_{ik} = 1$ then sample $\boldsymbol{x}_i$ belongs to cluster k. There can only be one 1 per column.

In effect this means finding an optimal distribution q among a family of distributions (Gaussian in our case). This optimization can be done by the introduction of a *variational distribution* of our choice, $q(\boldsymbol{\theta})$, to estimate the posterior $p(\boldsymbol{x}|\boldsymbol{\theta})$. In order to measure the similarity of q to p we would want to use the Kullback Leibler divergence,

$$KL\left(q(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{x})\right) = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})} d\boldsymbol{\theta} = \mathbb{E}_q \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{x})} \right]$$

However, the KL divergence of p from q gives rise to an expression containing the evidence, $p(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta}$, which unfortunately most oftenly is intractable. This is one of the reasons why we turn to VI in the first place.

However, we can navigate this issue by rewriting the KL-divergence and noticing that $p(\boldsymbol{x})$ is not a function of q:

$$\begin{aligned}\mathbb{E}_q \left[\log \frac{p(\boldsymbol{x}) q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \boldsymbol{x})} \right] =& \mathbb{E}_q \left[\log q(\boldsymbol{\theta}) \right] + \mathbb{E}_q \left[\log p(\boldsymbol{x}) \right] - \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{x}) \right] \\ =& \mathbb{E}_q \left[\log q(\boldsymbol{\theta}) \right] + \log p(\boldsymbol{x}) - \mathbb{E}_q \left[\log p(\boldsymbol{\theta}, \boldsymbol{x}) \right]\end{aligned}$$

We note here that $p(\boldsymbol{x})$ will be constant given the data $\boldsymbol{x}$. Thus, if we solve for $\log p(\boldsymbol{x})$ we will get an expression on the other side of the equality that necessarily must be constant. This term is:

$$\log p(\boldsymbol{x}) = \mathbb{E}_q\left[\log \frac{p(\boldsymbol{x})q(\boldsymbol{\theta})}{p(\boldsymbol{\theta},\boldsymbol{x})}\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\theta})\right] + \mathbb{E}_q\left[\log p(\boldsymbol{\theta},\boldsymbol{x})\right],$$

which we recognize as the KL-divergence plus an additional term, that we name *Evidence Lower Bound (ELBO)*. We thus have that

$$\log p(\boldsymbol{x}) = KL(q(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \boldsymbol{x})) + ELBO(q),$$

where, $ELBO(q) = \mathbb{E}_q\left[\log p(\boldsymbol{\theta},\boldsymbol{x})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\theta})\right]$.

We note that, since $\log p(\boldsymbol{x})$ is constant, minimizing the KL-divergence is equivalent to maximizing the $ELBO$-term (which is analytically tractable thanks to our choice of q!). We thus want to compute ELBO for the given model, which is given by expanding the joint distribution $p(\boldsymbol{\theta},\boldsymbol{x})$ into its independent parts, as well as using the mean field assumption, which states that $q(\boldsymbol{\theta}) = \prod_{i=1}^{k} q_i(\theta_i)$.

The ELBO for this model is thus given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) = &\sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} \mathbb{E}_q\left[\log p(\boldsymbol{c}_i)\right] \\ &+ \sum_{i=1}^{N} \mathbb{E}_q\left[\log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu})\right] - \sum_{k=1}^{K} \mathbb{E}_q\left[\log q(\boldsymbol{\mu}_k)\right] \\ &- \sum_{i=1}^{N} \mathbb{E}_q\left[\log q(\boldsymbol{c}_i)\right]\end{aligned} \tag{1}$$

2 Assignment 1

To find the ELBO we find the analytical expression for each of the terms making up the ELBO. For sake of brevity, the derivations for the first expression we will be relatively detailed, and later not as explicit with steps that are analogous to those performed for this first term.

Term 1 (The expected log prior over cluster centres)

Here we want to find

$$\sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\mu}_k)].$$

Here the subindex q for the expectation is with regards to $q(\boldsymbol{\mu}_k)$. We will use that:

$$p(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \boldsymbol{\alpha}, \sigma^2 \boldsymbol{I})$$

We begin by calculating the expectation and later apply the sum. By inserting the PDF for $\boldsymbol{\mu}_k$ under p we get that that

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\mu}_{\boldsymbol{k}})] =& \mathbb{E}_q \left[-\frac{p}{2} \left(\log 2\pi\sigma^2 \right) - \frac{1}{2\sigma^2} \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right) \right] \\ =& -\frac{p}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \mathbb{E}_q \left[\left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right) \right]\end{aligned}$$

Here we have used that everything except $\boldsymbol{\mu}_k$ is constant with respect to $q(\boldsymbol{\mu}_k)$, which we are taking expectation with respect to. This will again be very useful in coming derivations. We are also using that the prior is homoscedastic over all dimensions p, which simplifies the expression.

In the following we expand the term inside the brackets and use that

$$q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \boldsymbol{m}_k, s^2 k \boldsymbol{I}),$$

implying that

$$\begin{aligned}\mathbb{E}_q[\boldsymbol{\mu}_k] =& \boldsymbol{m}_k \\ \mathbb{E}_q[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k] =& Tr(s_k^2 \boldsymbol{I}) + \boldsymbol{m}_k^T \boldsymbol{m}_k \\ =& p s_k^2 + \boldsymbol{m}_k^T \boldsymbol{m}_k\end{aligned}$$

This result can be derived but was taken from [1]. We thus get that

$$\begin{aligned}\mathbb{E}_q \left[\left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right) \right] =& \mathbb{E}_q \left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k - 2 \boldsymbol{\mu}_k^T \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \boldsymbol{\alpha} \right] \\ =& \mathbb{E}_q \left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k \right] - 2 \mathbb{E}_q[\boldsymbol{\mu}_k^T \boldsymbol{\alpha}] + \mathbb{E}_q[\boldsymbol{\alpha}^T \boldsymbol{\alpha}] \\ =& \mathbb{E}_q \left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k \right] - 2 \mathbb{E}_q[\boldsymbol{\mu}_k^T] \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \boldsymbol{\alpha} \\ =& p s_k^2 + \boldsymbol{m}_k^T \boldsymbol{m}_k - 2 \boldsymbol{m}_k^T \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \boldsymbol{\alpha}\end{aligned}$$

By again closing this expression into a matrix multiplication we get the simplified expression

$$\mathbb{E}_q \left[\left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right)^T \left(\boldsymbol{\mu}_k - \boldsymbol{\alpha} \right) \right] = p s_k^2 + (\boldsymbol{m}_k - \boldsymbol{\alpha})^T (\boldsymbol{m}_k - \boldsymbol{\alpha})$$

Inserting this back into the original expression we get that

$$\mathbb{E}_q[\log p(\boldsymbol{\mu}_k)] = -\frac{p}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \left(p s_k^2 + (\boldsymbol{m}_k - \boldsymbol{\alpha})^T (\boldsymbol{m}_k - \boldsymbol{\alpha}) \right)$$

and thus that

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}_q[\log p(\boldsymbol{\mu}_k)] =&\sum_{k=1}^{K}-\frac{p}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}\left(ps_k^2+(\boldsymbol{m}_k-\boldsymbol{\alpha})^T(\boldsymbol{m}_k-\boldsymbol{\alpha})\right)\\ =&-K\frac{p}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}\sum_{k=1}^{K}\left(ps_k^2+(\boldsymbol{m}_k-\boldsymbol{\alpha})^T(\boldsymbol{m}_k-\boldsymbol{\alpha})\right)\end{aligned} \tag{2}$$

Term 2 (Expected log-prior over mixture assignments)

We want to find

$$\sum_{i=1}^{N}\mathbb{E}_q[\log p(\boldsymbol{c}_i)].$$

In this part we let the subindex q indicate $q(\boldsymbol{c}_i)$. Here we will use that

$$p(\boldsymbol{c}_i)=Cat(\boldsymbol{c}_i \mid \frac{1}{K},...,\frac{1}{K})=\frac{1}{K}$$

for all i (uniform prior over all cluster assignments). We thus get that

$$\begin{aligned}\sum_{i=1}^{N}\mathbb{E}_q[\log p(\boldsymbol{c}_i)] =&\sum_{i=1}^{N}\mathbb{E}_q[\log\frac{1}{K}]\\ =&-\sum_{i=1}^{N}\log K\\ =&-N\log K\end{aligned} \tag{3}$$

Term 3 (Expected log likelihood)

Here we will use that

$$\begin{aligned}p(\boldsymbol{x}_i \mid \boldsymbol{c}_i,\boldsymbol{\mu}) =&\mathcal{N}(\boldsymbol{x}_i \mid c_i^T\boldsymbol{\mu},\lambda^2\boldsymbol{I})\\ q(\boldsymbol{c}_i) =&Cat(\boldsymbol{c}_i \mid \boldsymbol{\phi}_i)\end{aligned}$$

where $\boldsymbol{\phi}_i=(\phi_{i,1},...,\phi_{i,K})$ contains the variational probabilities for each cluster $k=1,...,K$ to be selected for sample $\boldsymbol{x}_i$.

To facilitate the algebraic manipulations we can use that

$$p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu}) = \prod_{j=1}^{K} p(\boldsymbol{x}_i \mid \boldsymbol{\mu}_j)^{c_{ij}}$$

This gives that

$$\begin{aligned}\mathbb{E}_q[\log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu})] =& \mathbb{E}_q\left[\sum_{j=1}^{K} c_{ij} \log p(\boldsymbol{x}_i \mid \boldsymbol{\mu}_j)\right] \\ =& \sum_{j=1}^{K} \mathbb{E}_{q(c_i)}[c_{ij}] \mathbb{E}_{q(\boldsymbol{\mu}_j)}[\log p(\boldsymbol{x}_i \mid \boldsymbol{\mu}_j)]\end{aligned}$$

Here we use the mean field assumption, namely that $q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k) \prod_{i=1}^{N} q(\boldsymbol{c}_i)$, implying that $\boldsymbol{c}$ and $\boldsymbol{\mu}$ are independent. We can thus separate the terms into the product of two expectations. Furthermore we know that $\mathbb{E}_q[c_{ij}] = \phi_{ij}$. Continuing on, now with the subindex q denoting $q(\boldsymbol{\mu}_j)$ for sake of brevity, we thus have that

$$\begin{aligned}&\sum_{j=1}^{K} \mathbb{E}_{q(c_{ij})}[c_{ij}] \mathbb{E}_{q(\boldsymbol{\mu}_j)}[\log p(\boldsymbol{x}_i \mid \boldsymbol{\mu}_j)] \\ =& \sum_{j=1}^{K} \phi_{ij} \mathbb{E}_q\left[-\frac{p}{2} \log 2\pi\lambda^2 - \frac{1}{2\lambda^2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_j\right)^T \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j\right)\right]\end{aligned}$$

Similarly as with Term 1 we can move the expectation past constants, expand the matrix multiplication term and apply the distributive property of the expectation. We again apply the results for the moments of $\boldsymbol{\mu}_j$ from Term 1, and with analogous manipulations get that

$$\sum_{i=1}^{N} \mathbb{E}_q\left[\log p(\boldsymbol{x}_i \mid c_i, \boldsymbol{\mu})\right] = \sum_{i=1}^{N} \sum_{j=1}^{K} \phi_{ij} \left(-\frac{p}{2} \log 2\pi\lambda^2 - \frac{1}{2\lambda^2}\left(ps_j^2 + (\boldsymbol{x}_i - \boldsymbol{m}_j)^T (\boldsymbol{x}_i - \boldsymbol{m}_j)\right)\right) \tag{4}$$

Term 4 (Entropy of variational location posterior)

We want to find

$$\sum_{k=1}^{K} \mathbb{E}_q[\log q(\boldsymbol{\mu}_k)]$$

Here we use that

$$q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \boldsymbol{m}_k, s_k^2\boldsymbol{I})$$

We insert the density function into the expression and again let the expectation pass through all variables that are depending on $\boldsymbol{\mu}_k$ and get that

$$\begin{aligned}
&\mathbb{E}_q[\log q(\boldsymbol{\mu}_k)] \\
&\qquad =\mathbb{E}_q\left[-\frac{p}{2}\log 2\pi s_k^2 - \frac{1}{2s_k^2}(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T(\boldsymbol{\mu}_k - \boldsymbol{m}_k)\right] \\
&\qquad = -\frac{p}{2}\log 2\pi s_k^2 - \frac{1}{2s_k^2}\mathbb{E}_q\left[(\boldsymbol{\mu}_k - \boldsymbol{m}_k)^T(\boldsymbol{\mu}_k - \boldsymbol{m}_k)\right] \\
&\qquad = -\frac{p}{2}\log 2\pi s_k^2 - \frac{1}{2s_k^2}\left(\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right] - 2\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\right]\boldsymbol{m}_k + \boldsymbol{m}_k^T\boldsymbol{m}_k\right) \\
&\qquad = -\frac{p}{2}\log 2\pi s_k^2 - \frac{1}{2s_k^2}\left(ps_k^2 + \boldsymbol{m}_k^T\boldsymbol{m}_k - 2\boldsymbol{m}_k^T\boldsymbol{m}_k + \boldsymbol{m}_k^T\boldsymbol{m}_k\right) \\
&\qquad = -\frac{p}{2}\log 2\pi s_k^2 - \frac{1}{2s_k^2}ps_k^2 \\
&\qquad = -\frac{p}{2}\log 2\pi s_k^2 - \frac{p}{2}
\end{aligned}$$

Applying the sum yields

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}_q[\log q(\boldsymbol{\mu}_k)] \\
&\qquad =\sum_{k=1}^K -\frac{p}{2}\log 2\pi s_k^2 - \frac{p}{2} \\
&\qquad = -\frac{p}{2}\left(\sum_{k=1}^K \log 2\pi s_k^2\right) - \frac{Kp}{2}
\end{aligned} \tag{5}$$

Term 5 (Entropy of variational assignment parameter)

We want to find

$$\sum_{i=1}^N \mathbb{E}_q[\log q(c_i)].$$

Again, here the subindex q denotes expectation with respect to $q(\boldsymbol{\mu}_k)$. Here we use that

$$q(\boldsymbol{c}_i) = Cat(\boldsymbol{\phi}_i) = \prod_{j=1}^{K} \phi_{ij}^{\mathbb{1}\{c_{ij}=1\}}$$

Here we again use the relationship between expected value of an indicator variable and the probability of that variable.

$$\begin{aligned}
&\mathbb{E}_q[\log q(c_i)] \\
&\qquad = \mathbb{E}_q \left[\log \left(\prod_{j=1}^{K} \phi_{ij}^{\mathbb{1}\{c_{ij}=1\}} \right) \right] \\
&\qquad = \mathbb{E}_q \left[\sum_{j=1}^{K} \mathbb{1}\{c_{ij}=1\} \log \phi_{ij} \right] \\
&\qquad = \sum_{j=1}^{K} \mathbb{E}_q \left[\mathbb{1}\{c_{ij}=1\} \right] \log \phi_{ij} \\
&\qquad = \sum_{j=1}^{K} q(c_{ij}=1) \log \phi_{ij} \\
&\qquad = \sum_{j=1}^{K} \phi_{ij} \log \phi_{ij}.
\end{aligned}$$

This gives that

$$\sum_{i=1}^{N} \mathbb{E}_q[\log q(c_i)] = \sum_{i=1}^{N} \sum_{j=1}^{K} \phi_{ij} \log \phi_{ij} \tag{6}$$

Total expression for ELBO

Inserting equations (2) - (6) into (1) gives that the the ELBO can be evaluated by

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) = &\sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\mu}_k)] \\
&+ \sum_{i=1}^{N} \mathbb{E}_q \left[\log p(\boldsymbol{c}_i)\right] \\
&+ \sum_{i=1}^{N} \mathbb{E}_q \left[\log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu})\right] \\
&- \sum_{k=1}^{K} \mathbb{E}_q \left[\log q(\boldsymbol{\mu}_k)\right] \\
&- \sum_{i=1}^{N} \mathbb{E}_q \left[\log q(\boldsymbol{c}_i)\right] \\
= &-K\frac{p}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\left(ps_k^2 + (\boldsymbol{m}_k - \boldsymbol{\alpha})^T(\boldsymbol{m}_k - \boldsymbol{\alpha})\right) \\
&-N\log K \\
&+\sum_{i=1}^{N}\sum_{j=1}^{K}\phi_{ij}\left(-\frac{p}{2}\log 2\pi\lambda^2 - \frac{1}{2\lambda^2}\left(ps_j^2 + (\boldsymbol{x}_i - \boldsymbol{m}_j)^T(\boldsymbol{x}_i - \boldsymbol{m}_j)\right)\right) \\
&+\frac{p}{2}\left(\sum_{k=1}^{K}\log 2\pi s_k^2\right) + \frac{Kp}{2} \\
&-\sum_{j=1}^{K}\phi_{ij}\log\phi_{ij} \qquad (7)
\end{aligned}
$$

2.1 Assignment 2

Finding $\phi_{i,k}$

We now want to show that the variational update for the cluster assignment of the ith sample to the kth cluster is

$$\phi_{i,k} \propto \exp\left\{\frac{\boldsymbol{x}_i^T\mathbb{E}[\boldsymbol{\mu}_k]}{\lambda^2} - \frac{\mathbb{E}[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k]}{2\lambda^2}\right\} \qquad (8)$$

In order to find the expression we need to show that $q^*(c_{ij})$ is proportional to (8). The optimal parameter is found by taking expectation of the log joint distribution with respect to all latent variational parameters except the one in question, i.e. with respect to $\boldsymbol{\mu}$ and $\boldsymbol{c}_{-i}$. Here the notation $\boldsymbol{c}_{-i}$ indicates all variables in $\boldsymbol{c}$ except $\boldsymbol{c}_i$.

We begin by rewriting the log joint so that it only contains the relevant variables:

$$\begin{aligned}\log p(\boldsymbol{x}, \boldsymbol{c}_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu}) = & \log p(\boldsymbol{\mu}) + \sum_{j \neq i} \left(\log p(\boldsymbol{c_j}) + \log p(\boldsymbol{x}_j \mid \boldsymbol{c}_j, \boldsymbol{\mu})\right) \\ & + \log p(\boldsymbol{c_i}) + \log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu})\end{aligned}$$

We want to find the optimal parameter for $\boldsymbol{c}_i$ and thus the maximization will be taken only with respect to variables that are functions of $\boldsymbol{c}_i$.

$$\begin{aligned}q^*(\boldsymbol{c}_i) & = \exp\left\{\mathbb{E}_{q(\boldsymbol{c}_{-i}, \boldsymbol{\mu})}\left[\log p(\boldsymbol{x}, \boldsymbol{c}_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu})\right]\right\} \\ & \propto \exp\left\{\mathbb{E}_{q(\boldsymbol{c}_{-i}, \boldsymbol{\mu})}[\log p(\boldsymbol{c}_i + \log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu}]\right\} \\ & = \exp\left\{\log p(\boldsymbol{c}_i) + \mathbb{E}_{q(\boldsymbol{c}_{-i}, \boldsymbol{\mu})}[\log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu}]\right\}\end{aligned}$$

Above we used that the expectation is taken with respect to all variational parameters except $\boldsymbol{c}_i$, and thus the expectation passes through it unchanged. We now follow the hint that $p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k)^{c_{i,k}}$. This gives that

$$\begin{aligned}& \exp\left\{\log p(\boldsymbol{c}_i) + \mathbb{E}_q[\log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu}]\right\} \\ & = \exp\left\{\log p(\boldsymbol{c}_i) + \mathbb{E}_{q(\boldsymbol{c}_{-i}, \boldsymbol{\mu})}\left[\sum_{k=1}^{K} c_{i,k} \log p(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k)\right]\right\} \\ & = \exp\left\{\frac{1}{K} + \sum_{k=1}^{K} c_{i,k} \mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[C - \frac{1}{2\lambda^2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)^T\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)\right]\right\}\end{aligned}$$

In the following calculations we will consider everything that is not a function of $\boldsymbol{c}_i$ as constants, and these will be disposed of under the sign of proportionality. This gives that

$$\begin{aligned}& \propto \exp\left\{\sum_{k=1}^{K} c_{i,k}\left(-\frac{1}{2\lambda^2}\mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)^T\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k\right)\right]\right)\right\} \\ & = \exp\left\{\sum_{k=1}^{K} c_{i,k}\left(-\frac{1}{2\lambda^2}\left(\mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[\boldsymbol{x}_i^T \boldsymbol{x}_i\right] - 2\mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[\boldsymbol{\mu}_k\right]\boldsymbol{x}_i + \mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k\right]\right)\right)\right\} \\ & \propto \exp\left\{\sum_{k=1}^{K} c_{i,k}\left(\frac{1}{\lambda^2}\mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[\boldsymbol{\mu}_k\right]\boldsymbol{x}_i - \frac{1}{2\lambda^2}\mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k\right]\right)\right\}\end{aligned}$$

Now considering only the kth entry for $\phi_{i,k}$ we finally get that

$$q^*(c_{i,k}) = \phi_{i,k} \propto \exp\left\{\frac{1}{\lambda^2}\mathbb{E}_{q(\boldsymbol{\mu}_k)}\left[\boldsymbol{\mu}_k\right]\boldsymbol{x}_i - \frac{1}{2\lambda^2}\mathbb{E}_{q(\boldsymbol{\mu}_k^T)}\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]\right\},$$

Which we wanted to prove.

This can also be done by differentiating ELBO, (eq. (7)) w.r.t. $\phi_{i,k}$ and solving for 0. We thus find the value for which $\phi_{i,k}$ maximizes the ELBO and thus minimizes the KL-divergence between q and p. This becomes quite handy as there are only two terms that depend on $\phi_{i,k}$. From (7) we get that:

$$\begin{aligned}
&\frac{\partial}{\partial\phi_{i,k}}ELBO(\boldsymbol{m}, \boldsymbol{s^2}, \boldsymbol{\Phi}) \\
=&\frac{\partial}{\partial\phi_{i,k}}\bigg(\phi_{i,k}\big(-\frac{p}{2}\log 2\pi\lambda^2 - \frac{1}{2\lambda^2}\mathbb{E}_q\big[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\big]\big) \\
&-\phi_{i,k}\log\phi_{i,k} + const\bigg) \\
=&-\frac{p}{2}\log 2\pi\lambda^2 - \frac{1}{2\lambda^2}\mathbb{E}_q\big[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\big] - (\log\phi_{i,k} + 1) \\
\propto&-\frac{1}{2\lambda^2}\left(\mathbb{E}_q\left[\boldsymbol{x}_i^T\boldsymbol{x}_i\right] - 2\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\right]\boldsymbol{x}_i + \mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right]\right) \\
\propto&\frac{1}{\lambda^2}\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\right]\boldsymbol{x}_i - \frac{1}{2\lambda^2}\mathbb{E}_q\left[\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right], \qquad (9)
\end{aligned}$$

which, again, is what we wanted to show.

Finding $\boldsymbol{m}_k$ and s_k^2

We are given that

$$q^*(\boldsymbol{\mu}_k) \propto \exp\left\{\log p(\boldsymbol{\mu}_k) + \sum_{i=1}^{N}\mathbb{E}_{q(\boldsymbol{\mu}_{-k},\boldsymbol{c}_i)}\left[\log p(\boldsymbol{x}_i \mid \boldsymbol{c}_i, \boldsymbol{\mu})\right]\right\}$$

and that the expression is proportional to a product of Gaussians, wherefore the resulting distribution is also a Gaussian. We are given the exponent, which we wish to rewrite (complete the squares) to be able to identify the mean vector and covariance matrix of the distribution (since the first and second moment uniquely determine the distribution).

We do this by grouping together terms that are linear and quadratic in terms of $\boldsymbol{\mu}_k$, and then compare these to the expected structure of our resulting distribution. We can thus disregard of terms that are constant w.r.t. $\boldsymbol{\mu}_k$.

$$\begin{aligned}
-\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} &+ \sum_{i=1}^N \phi_{i,k}\left(-\frac{1}{2\lambda^2}\left(\boldsymbol{x}_i^T\boldsymbol{x}_i - 2\boldsymbol{\mu}_k^T\boldsymbol{x}_i + \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\right)\right) \\
&\propto -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^N \phi_{i,k}\boldsymbol{\mu}_k^T\boldsymbol{x}_i - \frac{1}{2\lambda^2}\sum_{i=1}^N \phi_{i,k}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k \\
&= -\frac{\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k}{2\sigma^2} + \boldsymbol{\mu}_k^T\boldsymbol{x}_i\frac{1}{\lambda^2}\sum_{i=1}^N \phi_{i,k} - \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\frac{1}{2\lambda^2}\sum_{i=1}^N \phi_{i,k} \\
&= -\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\left(\frac{1}{\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^N \phi_{i,k}\right) + \boldsymbol{\mu}_k^T\frac{1}{\lambda^2}\sum_{i=1}^N \phi_{i,k}\boldsymbol{x}_i \qquad (10)
\end{aligned}$$

We now compare (10) to the expected structure of the resulting distribution, which will have the form

$$\begin{aligned}
\boldsymbol{\mu}_k &\sim \mathcal{N}(\boldsymbol{m}_k, s_k^2\boldsymbol{I}) \\
p(\boldsymbol{\mu}_k) &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k\left(\frac{1}{s_k^2}\right) + \boldsymbol{\mu}_k^T\frac{1}{s_k^2}\boldsymbol{m}_k\right\} \qquad (11)
\end{aligned}$$

By matching the coefficients of the quadratic terms w.r.t. $\boldsymbol{\mu}_k$ in (10) and the exponent of (11) we can now determine that

$$\frac{1}{s_k^2} = \frac{1}{\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^N \phi_{i,k},$$

which after solving for s_k^2 gives

$$s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\lambda^2}\sum_{i=1}^N \phi_{i,k}} \qquad (12)$$

Similarly we can by matching the linear terms of (10) and the exponent of (11) determine that

$$\frac{1}{s_k^2}\boldsymbol{m}_k = \frac{1}{\lambda^2}\sum_{i=1}^N \phi_{i,k}\boldsymbol{x}_i,$$

giving that

$$\boldsymbol{m}_k = \frac{s_k^2}{\lambda^2}\sum_{i=1}^N \phi_{i,k}\boldsymbol{x}_i \qquad (13)$$

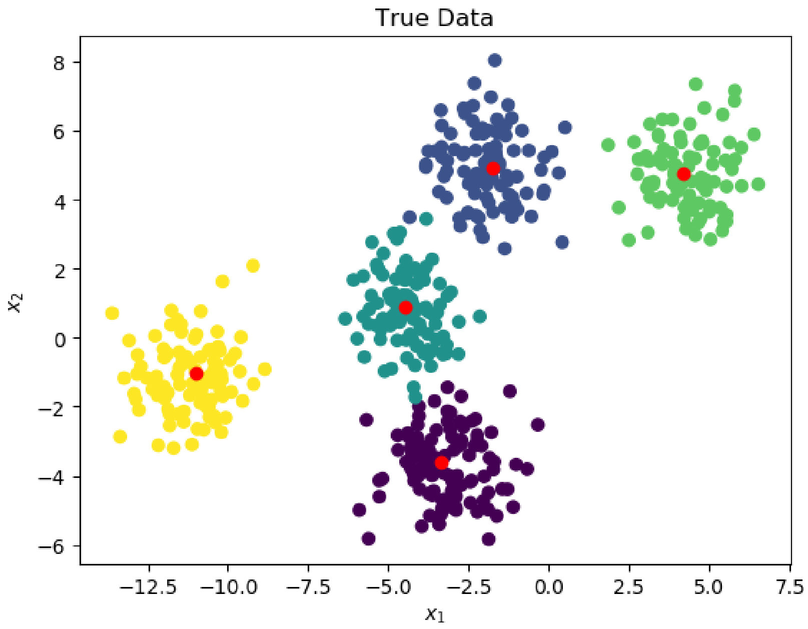

Figure 1: Visualization of the data, the cluster assignments and the true underlying cluster centres.

2.2 Assignment 4

By applying these update equations we can run an iterative scheme, in which we randomly initialize the variational parameters, $\boldsymbol{\mu}$, $\boldsymbol{s}^2$ and $\boldsymbol{\phi}$. We then in turn update ϕ_i for all $i = 1, ..., N$, based on the initialization of $\boldsymbol{\mu}$ and $\boldsymbol{s}$ through (9) (and each column ϕ_j is normalized by $\sum_{i=1}^{N} \phi_{ij}$). This result is then used to compute the updates for the new $\boldsymbol{\mu}$ and $\boldsymbol{s}^2$ in turn, by using (12) and (13).

The result in Figure 2 shows the best result after running the algorithm 10 times and picking the best result. We do this since the EM-like algorithms are prone to getting stuck in local minimum, so by initializing the algorithm with different vectors/matrices we obtain multiple solutions and pick the one with the best (highest) ELBO. We find that the algorithm accurately manages to capture all cluster assignments and locations. Only one data point is classified erroneously. In Figure 3 we see that the ELBO monotonically increases and converges after 10 iterations.

References

[1] https://www.math.uwaterloo.ca/ hwolkowi/matrixcookbook.pdf

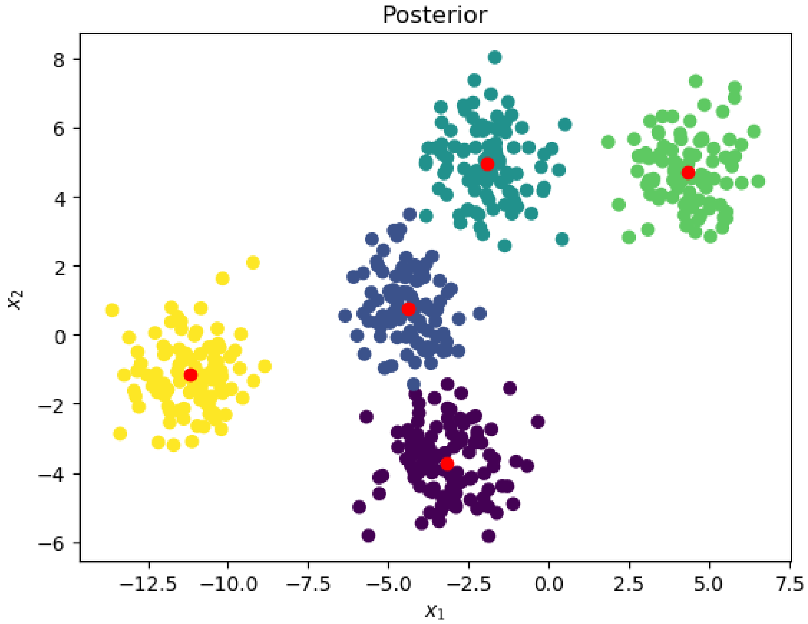

Figure 2: Visualization of the data, the inferred cluster assignments and the inferred underlying cluster centres, as estimated by the Coordinate Ascent VI algorithm.

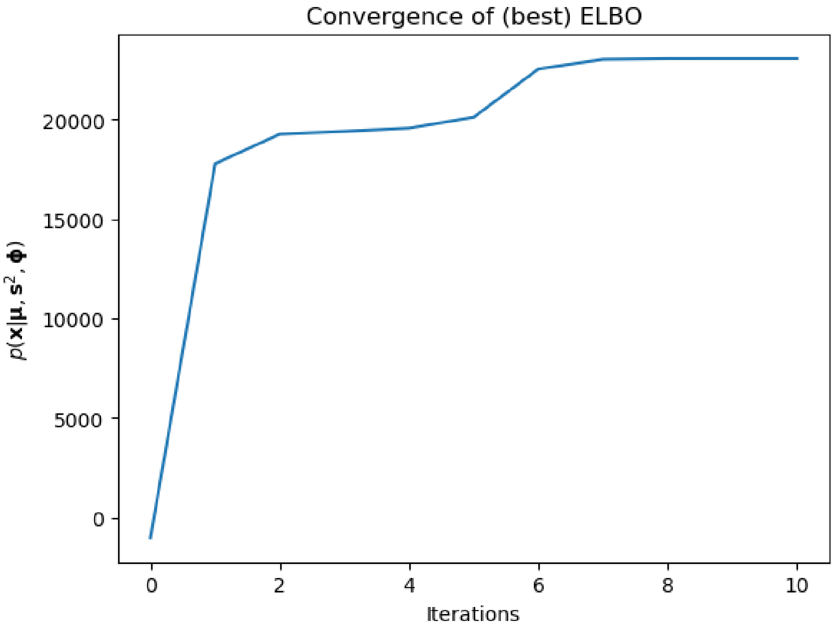

Figure 3: Plot of the Estimated Lower Bound (ELBO) for the resulting inference in figure 2.